Chemometrics and Intelligent Laboratory Systems 104 (2010) 83–94

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

A generic linked-mode decomposition model for data fusion

Iven Van Mechelen [a,*], Age K. Smilde [b]

[a] Research Group on Quantitative Psychology and Centre for Computational Systems Biology (SymBioSys), KU Leuven, Tiensestraat 102-box 3713, Leuven, Belgium
[b] Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV, Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 2 November 2009
Received in revised form 12 April 2010
Accepted 17 April 2010
Available online 27 April 2010

Keywords:
Data fusion
Multiblock data
Multiset data
Functional genomics

ABSTRACT

As a consequence of our information society, not only more and larger data sets become available, but also data sets that include multiple sorts of information regarding the same system. Such data sets can be denoted by the terms coupled, linked, or multiset data, and the associated data analysis can be denoted by the term data fusion. In this paper, we first give a formal description of coupled data, which allows the data-analyst to typify the structure of a coupled data set at hand. Second, we list two meta-questions and a series of complicating factors that may be useful to focus the initial content-driven research questions that go with coupled data, and to choose a suitable data-analytic method. Third, we propose a generic framework for a family of decomposition-based models pertaining to an important subset of data fusion problems. This framework is intended to constitute both a means to arrive at a better understanding of the features and the interrelations of the specific models subsumed by it, and as a powerful device for the development of novel, custom-made data fusion models. We conclude the paper by showing how the proposed formal data description, meta-questions, and generic model may assist the data-analyst in choosing and developing suitable strategies for the treatment of coupled data in practice. Throughout the paper we illustrate with examples from the domain of systems biology.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

One of the dominant features of systems biology is the explosion of data. It is more rule than exception that multiple sets of data are collected pertaining to the same biological system. Analyzing all such data sets simultaneously permits a global view on the biological system under study and, hence, attracts increasingly attention. Such an endeavor goes under different names, including *data fusion* [1,2], *analysis of coupled or linked data* [3], *multiset* or *multiblock data analysis* [4], and *integrative data analysis* [5]. This terminology is diffuse: Data fusion is not clearly defined and integrative data analysis does not necessarily coincide with multiblock data analysis in all its applications. We will use the term data fusion throughout and give a clear definition later (see Section 4).

Data fusion implies a major challenge for data-analysts. There are at least three reasons for this: (1) It involves very complex data, the structure of which is not always too easy to grasp, (2) it goes with a very broad range of research questions, with different possible questions being associated with the same data set, and (3) quite a few fairly different data-analytic methods are available to address data fusion problems and many others still need to be developed.

In the present paper we will offer the data-analyst a hold to deal with this utmost complex and challenging situation. More specifically: (1) We will give a formal definition of coupled data, which will include several subtypes of such data, and which will allow the data-analyst to typify the structure of a coupled data set at hand (Section 2), (2) we will list two meta-questions and a series of complicating factors that may act as useful tools for the data-analyst to focus the initial content-driven research questions, and as beacons for the subsequent choice of a suitable data-analytic method (Section 3), (3) we will propose a generic framework for a family of decomposition-based models pertaining to an important subset of data fusion problems. This framework is intended to constitute both a means to arrive at a better understanding of the features and the interrelations of the specific models subsumed by it (which may be most helpful in the choice of a suitable method to analyze a coupled data at hand), and a powerful device for the development of novel, custom-made data fusion models (Section 4). We will conclude this paper by showing how our formal data description, meta-questions, and generic model may assist the data-analyst in choosing (resp. developing) suitable strategies for the treatment of a coupled data set at hand (Section 5).

This paper defies categorization in several respects: (a) It is not intended as a review of earlier data fusion work. Rather, on the one hand, it will present a novel framework to understand coupled data structures and to focus associated research questions; on the other hand, it will introduce a novel generic decomposition model that subsumes an important subset of specific data fusion methods. However, to do so, the paper will start from a number of existing data fusion studies and from a few existing concepts; also, the paper

* Corresponding author.
E-mail address: Iven.VanMechelen@psy.kuleuven.be (I. Van Mechelen).

0169-7439/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2010.04.012

will show how several existing data fusion methods are subsumed by the proposed generic model. (b) The paper will propose a novel framework for the description and analysis of coupled data, and, as such, it is not a mere tutorial. However, an attempt will be made to explain the rather complex data fusion setting and associated models as clearly as possible and in a didactic way. (c) The argument of the paper will be fairly abstract and theoretical, and results of specific data fusion analyses of particular data sets will not be reported. Yet, throughout the paper, ample reference will be made to well-defined (hypothetical as well as real) data sets, and an attempt will be made to show how the framework as outlined in the paper may act as a guiding tool in data-analytic practice.

2. Coupled data

2.1. Examples of data structures

In metabolomics it is increasingly common to measure the same set of samples on different analytical platforms to obtain a comprehensive view of the metabolites in those samples [2,6]. One can also study functional genomics measurements of the same type performed in different organisms [7], or in different compartments of the same organism, for example, in plasma and tissue [8]. Data can further be obtained of the same organism in terms of gene expression, ChIP-on-Chip, alternative splicing (Exon arrays), copy-number measurements (CGH arrays) and polymorphism genotyping (SNP arrays) [9]; this may stretch even further by also including text mining results [1].

The references above suggest that data fusion problems are abundant in systems biology. Moreover, they also illustrate that data fusion problems and the associated data can take a diversity of structural forms and are not easily categorized. As a starting point for our attempt to deal with this challenge, we will pick out three particular data sets as guiding examples.

Guiding example 1 stems from a microbial metabolomics study on several *Escherichia coli* strains that were cultivated under different environmental conditions [2]. The metabolomes of these fermentations were analyzed using two different measurement platforms, LC–MS and GC–MS. This resulted in two fermentation by metabolite data matrices pertaining to the same set of fermentations.

Guiding example 2 stems from a study by Ref. [1] on gene prioritization. Data from this study pertained to a set of genes that comprised both a subset of training genes that were known to be associated with some disease, and a subset of test genes. On these genes, information was available from a broad range of data sources, including occurrence in abstracts (in EntrezGene), functional annotation information (Gene Ontology), transcriptomics information, and data on transcriptional motifs.

Guiding example 3 is taken from a study [10] on yeast cell cycle time courses. In this study, for cultures put under different oxidative stress conditions, mRNA expression levels were measured for 4329 yeast *Saccharomyces cerevisiae* genes at 13 synchronized time points. In addition, also for the very same genes, protein binding information was available for a number of transcription factors.

2.2. Formal characterization of coupled data

In order to typify the different possible data structures we will rely on a conceptual framework introduced by Ref. [11]. The basic constituents of coupled data are *data blocks*. Data blocks can be considered a mapping **B** from a Cartesian product $S = S_1 \times S_2 \times \ldots \times S_N$ to some (typically univariate) range Y: For each N-tuple $(s_1, s_2, \ldots, s_N)$ with $s_1 \in S_1, s_2 \in S_2, \ldots, s_N \in S_N$ a value $\mathbf{B}(s_1, s_2, \ldots, s_N)$ from Y is recorded. The number of sets in the Cartesian product of the domain is called the number of *ways* and the number of distinct sets in the same product is called the number of *modes* in the data. As an example, one may look at the three single data blocks as depicted in Fig. 1. Panel (a) of this figure pertains to gene by transcription factor binding information; as such it is an example of two-way two-mode (gene by transcription factor) data. Panel (b) pertains to longitudinal transcriptomics information with regard to a number of tissues; its structure is gene by time by tissue, and therefore this data block is three-way three-mode. Finally, panel (c) pertains to similarity information between all possible pairs of genes as derived from different databases; the structure of this data block is gene by gene by database, and therefore it can be characterized as three-way two-mode.

As an aside, one may note that some data collection procedures yield data blocks for which some parts are structurally missing. As an example one may consider a longitudinal transcriptomics data collection procedure, which is similar to the one as illustrated by panel (b) of Fig. 1, except for the fact that the measurements for the different tissues are no longer taken at comparable time points. This comes down to a data structure for which time points are nested within tissues, rather than that the gene and time point modes are

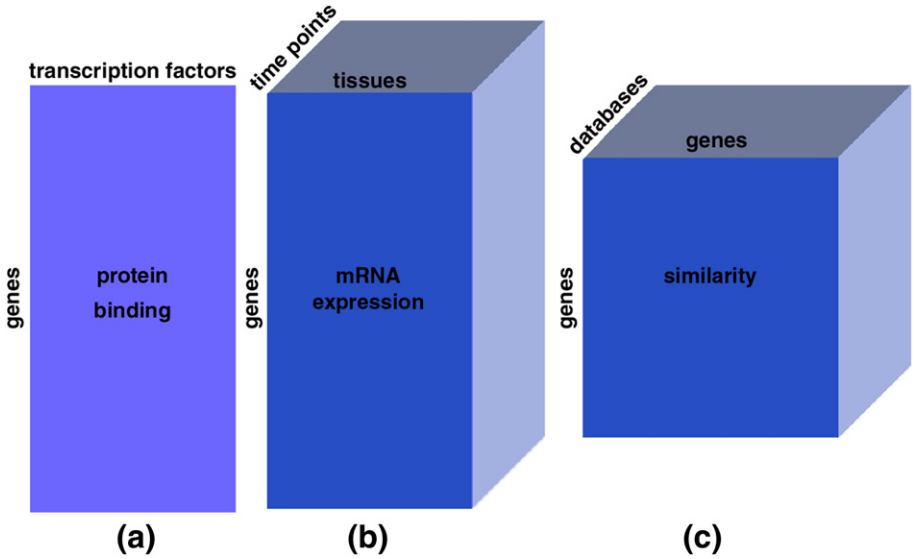

Fig. 1. Examples of three types of single data blocks: (a) two-way two-mode gene by transcription factor data, (b) three-way three-mode gene by time point by tissue data, (c) three-way two-mode gene by gene by database similarity data.

fully crossed. Such a data structure could be formalized as a mapping from a gene by tissue by time point Cartesian product, with the time point mode comprising the time points for all tissues, and by structural missingness of the measurements for each tissue at the time points of all other tissues (see also Ref. [12]).

Making use of the concepts of ways, modes, and data blocks, coupled data can now be defined as a connected collection of data blocks, with the connections between blocks consisting of shared modes. To illustrate, we revisit the three guiding examples that we introduced in the previous section. They give rise to relatively simple coupled data structures, which are graphically represented in Fig. 2.

The first guiding example pertained to the microbial metabolomics data set as analyzed by Ref. [2]. This can be considered a case of two coupled two-way two-mode data blocks that are connected through a common fermentation mode. This data structure is graphically represented by panel (a) of Fig. 2. The second guiding example stems from the study by Ref. [1] on gene prioritization. It yields a data structure that consists of ten two-way, two-mode and one-mode data blocks, which all share a common gene mode. Part of this structure is graphically represented by panel (b) of Fig. 2. One may note the fan-like nature of the representation, which immediately visualizes the common mode in terms of the shared side of the data rectangles. The third guiding example is taken from the study from Ref. [10]. It involves a three-way three-mode (gene by stress condition by time point) and a two-way two-mode gene by transcription factor block, which share the gene mode. This structure is graphically represented in panel (c) of Fig. 2.

It may be useful to emphasize that coupled data structures can assume more complex forms than those of Fig. 2. Graphical representations of examples of three more complex coupled data structures can be found in Fig. 3.

These examples illustrate three types of complexity. First, coupled data can consist of three or more connected data blocks not all pairs of which share at least one common mode. As an example, imagine that the data from the study of Ref. [10] would be supplemented by background information on the stress conditions in the form of a stress condition by feature data block. This would yield the coupled data structure as depicted in panel (a) of Fig. 3. This can easily be generalized to coupled data structures with an even more complex jigsaw puzzle pattern. Secondly, two data blocks may have more than a single mode in common. As an example, consider the data that are schematically represented in panel (b) of Fig. 3. Those pertain to (synchronized) longitudinal transcriptomics data with regard a panel of tissues on the one hand, and with regard to a set of stem cells on the other hand. Such data could be considered to comprise two data blocks, a first one pertaining to the tissues and a second one pertaining to the stem cells. Those two data blocks are both three-way three-mode, and they have two modes in common: the gene mode and the time point mode. Thirdly, up to now we have only considered data blocks that are fully coupled in that they fully share one or more modes. Consider, however, the data structure as represented in panel (c) of Fig. 3. This pertains to a comparative genomics study with transcriptomics data collected from two different microbial organisms (e.g., *E. coli* and *Bacillus subtilis*), that each was cultivated under a number of experimental conditions. This yields two two-way two-mode data blocks, with as common mode as the gene mode. This gene mode, however, is now only partially shared, with only the orthologous genes of the two organisms being involved in the linkage.

3. Research questions

3.1. Examples of research questions

Coupled data may go with a plethora of questions. To illustrate, we revisit once again our three guiding examples. The first of those pertained to the microbial metabolomics study by Ref. [2] with metabolomes of *E. coli* fermentations that were analyzed using two different measurement platforms, LC–MS and GC–MS. A key question in this case is which are the common and distinctive aspects of the metabolome that are captured by the two measurement platforms.

The second guiding example was taken from a study by Ref. [1] on gene prioritization. Starting from a set of training genes that are known to be associated with some disease, these authors wanted to identify the most promising test genes that might be also involved in the disease in question. For this purpose, they could rely on a broad range of data sources. Stated otherwise, in this case the research question can be summarized as 'Which test genes are most similar to the training genes across the whole of all data blocks?'

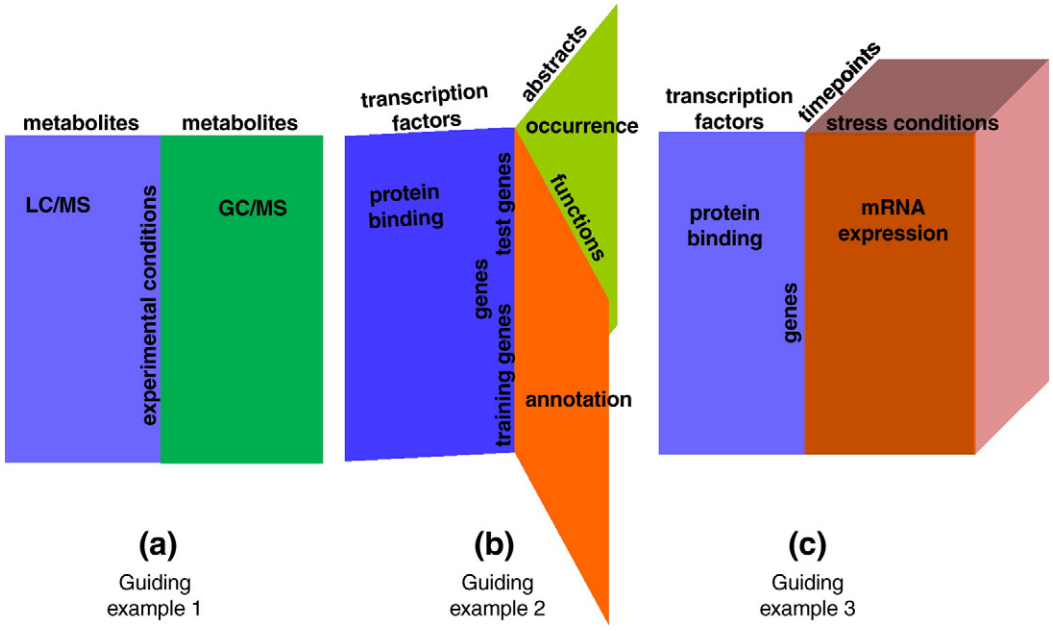

Fig. 2. Three relatively simple coupled data sets corresponding to three guiding examples, taken from real studies: (a) microbial metabolomics data as analyzed by Ref. [2], (b) data from study by Ref. [1] on gene prioritization, (c) data on yeast cell cycle time course from study by Ref. [10].

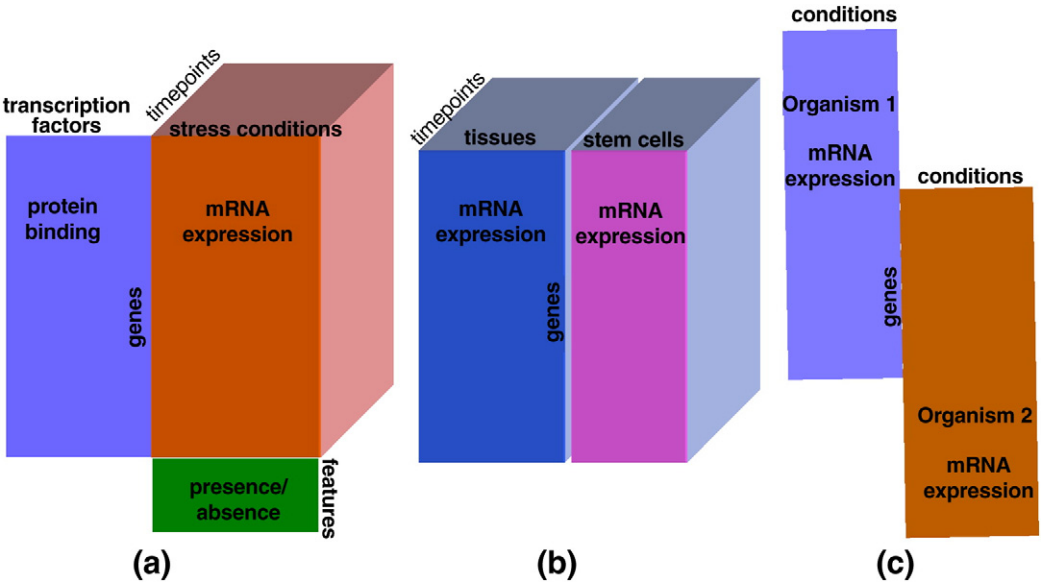

Fig. 3. Three examples of complex coupled data: (a) coupled data blocks not all pairs of which share at least one common mode, (b) two data blocks that have more than a single mode in common, (c) two data blocks with a common mode that is partially shared only.

The third guiding example was taken from the study by Ref. [10] on yeast cell cycle time courses, with a three-way three-mode longitudinal yeast transcriptomics block that was coupled to a two-way two-mode protein binding block through the gene mode. An important underlying scientific question in this case reads whether transcription binding predicts gene expression profiles across time.

3.2. Meta-questions

When analyzing the questions above as well as similar questions on coupled data, one may typify them in terms of a number of generic characteristics. Those characteristics can be linked to two meta-questions. In order to arrive at a suitable data-analytic strategy, it may be of utmost importance to carefully look for an answer to these.

The first meta-question pertains to which information is to be derived from the data. This meta-question further comprises two parts. The first part concerns which information is to be derived from each data block. One possible option in this regard could be that one may be interested in the full information as included in the data block. This means that one may wish to have a model that accounts for the actual entries as included in the data block, or for a data-analytic approach that allows the researcher to reconstruct those entries as closely as possible. For instance, in the microbial metabolomics data as studied by Ref. [2] (see Guiding example 1 above), one might wish to capture for each measurement platform the full metabolome of each fermentation under study. As an alternative option, one may only be interested in some partial information as implied by the target data block. Examples of such partial information include similarity information between the elements of a particular data mode, or merely the interaction or dependence information between two or more modes as implied by a data block. To illustrate, we return to the paper by Ref. [1] on gene prioritization (see Guiding example 2 above). Looking at one of the blocks (i.e., matrices) as included in the data of this study, the primary focus of the researchers is not on the full information in the block, but only on between-gene similarity information that may be derived from the block.

The second part of the first meta-question pertains to which information is to be derived from the whole of the data blocks. One possible alternative at this point could be that the primary research interest resides in consensus information that may be derived from the data blocks through some voting or averaging procedure. For instance, in Guiding example 2 above on gene prioritization, the research focus was on a consensus ranking of the test genes with respect to similarity with the training genes that were known to be associated with the target disease. As a second alternative, one could be interested in both commonalities and differences between the different data blocks under study. For instance, in the microbial metabolomics case of Guiding example 1, the interest was in common as well as distinctive aspects of the metabolome as captured by the LC/MS and GC/MS measurement platforms. As a third alternative, research interest could focus on the linkage or linking relations between coupled data blocks. For instance, in the study of Guiding example 3 on yeast cell cycle time courses, the research focus was on the linking relation between the protein binding and the gene expression blocks.

The second meta-question pertains to the roles of the different coupled data blocks in the overall data analysis. This meta-question, too, comprises two parts. The first pertains to whether the different data blocks do assume qualitatively different roles in the overall analysis; if this would not be the case, the roles of the distinct blocks can be called exchangeable. A case in which exchangeability does not hold is a prediction situation in which a first block is considered a block of predictor information and a second block a criterion. This is exemplified by Guiding example 3 on yeast cell time course, in which the protein binding block could be assigned the role of predictor, whereas the longitudinal transcriptomics block could be considered the criterion to be predicted.

The second part of the second meta-question is more quantitative in nature: It pertains to whether the different data blocks have equal or different levels of priority or importance. For instance, when dealing with the data of Guiding example 3, unlike in the paper by Ref. [10], the primary focus could be on an in-depth understanding of the expression profiles, with the protein binding information being only of secondary or minor research interest.

3.3. Complicating factors

In problems of data fusion, a number of complicating factors may show up. Those have to be dealt with in an appropriate way, to allow for a meaningful data analysis. Below we will discuss four such complicating factors in somewhat more detail.

A first complicating factor pertains to the links between the different coupled data blocks involved. In quite a few cases, such links may imply an alignment problem. (Note that a number of authors use

the term data fusion exclusively to denote this alignment problem.) As an example, one may think of transcriptomics data as collected from two microbial organisms, each organism being measured under a number of experimental conditions. In dealing with the resulting coupled data matrices, one may wonder which genes in the leftmost matrix from Organism 1 are orthologous to which genes in the rightmost matrix from Organism 2 (see also panel (a) of Fig. 4). Similarly, in a different research area (viz., that of brain imaging studies), one may consider fMRI data as collected from two or more different persons (see also panel (b) of Fig. 4). When dealing with such data, one may wonder which voxel from the brain of the first person corresponds to which voxel from the brain of the second. When dealing with coupled data, this and similar alignment problems are ubiquitous, and in general far from trivial to solve. One possible way to deal with them, is through the construction of some suitable mappings (e.g., a mapping of voxels onto a so-called 'standard brain', as is typically included in most preprocessing procedures for brain imaging data).

A second complicating factor pertains to comparability of different data entries. To clarify this complicating factor, one may consider data pertaining to a set of patients from whom different measurements have been taken (blood pressure, heart rate, lung capacity, etc.), which results in a two-way two-mode patient by variable matrix. When looking at the entries of this matrix, it should be clear that it is difficult, if not meaningless, to compare a data entry pertaining to blood pressure with a data entry pertaining to lung capacity (either from the same or from two different patients). Such a lack of comparability can be considered to reflect the fact that blood pressure and lung capacity are expressed on different measurement scales (which can also be denoted by the term 'lack of commensurability'). Lack of comparability/commensurability may imply a major obstacle for an appropriate data analysis, as comparability/commensurability is implicitly required for most data-analytic methods. To get a better intuitive idea about this, one may simply consider data-analytic methods that rely on a least squares estimation procedure. The loss function that is to be optimized by such methods typically involves a sum of squared residuals across the entire data set (which implies the tacit assumption that the residuals in question, and therefore also the corresponding data entries, are commensurable, indeed). Comparability/commensurability bears a close relation to preprocessing, in that in a number of cases one may hope to rectify some lack of comparability/commensurability through a suitable preprocessing procedure (e.g., standardization). Yet, whether such procedures do restore comparability, indeed, is a far from trivial issue. Within a data fusion context, comparability/commensurability is to be taken care of within each data block as well as between the different data blocks. The latter is especially important if one considers global models for coupled data at hand, along with global objective or loss functions that are to be optimized in the associated data analysis (see below).

The third and fourth complicating factors pertain to possible heterogeneities of the different data blocks, which may hamper the data fusion process. To understand this better, one should note that all forms of data fusion involve some form of aggregation of the different data blocks, which may be considered a kind of voting procedure, either in the narrow or in the broad sense. One may wish this voting procedure to take place in an optimal way, optimality to be considered here as leading to the best possible inferences about some true structural aspects underlying the data.

In particular, the third complicating factor pertains to possible differences between the data blocks in size. It should be clear that in systems biology research, differences in size show up rather frequently, for instance because some data blocks (unlike others) may include very large modes (e.g., a set of genes). Obvious differences in size also show up in multiset data with data blocks that differ in the number of ways involved. For instance, in the example of panel (c) of Fig. 2, the three-way transcriptomics block is obviously much larger than the two-way protein binding block linked to it.

The fourth complicating factor pertains to possible differences between the data blocks in noise characteristics. Beyond issues of comparability as already touched upon above, data blocks obviously may also be susceptible to different amounts and types of error, for example, because they have been obtained through quite different measurement procedures. For instance, one may safely conjecture that the linked data matrices of panel (b) of Fig. 2 differ quite considerably in their implied noise levels. To optimize subsequent inferences, one may obviously wish to take such noise heterogeneity into account properly.

4. Generic model

Data fusion can be broadly defined as any type of data-analytic or statistical procedure that involves coupled data as defined above. Methods of data fusion can further focus on any possible set of research questions on coupled data that may be typified by answers to the meta-questions as outlined above, and/or on any set of complicating factors.

In the present section of this paper, we will focus on one important subclass of data fusion problems. This subclass will not be restricted in terms of data structures. Yet, it will be restricted in terms of the research questions it addresses. In particular, in terms of the first meta-question, we will exclusively focus on problems with the aim of

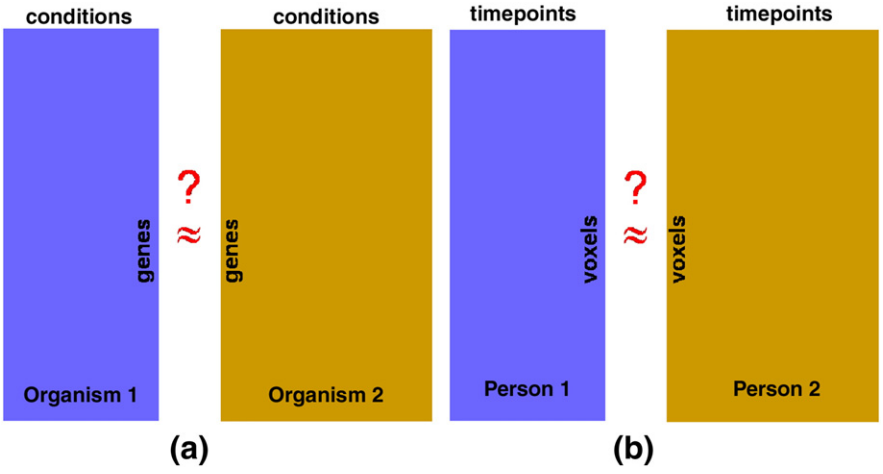

Fig. 4. Two examples of alignment/mapping problems: (a) coupled transcriptomics data stemming from two different microbial organisms, (b) coupled fMRI data stemming from two different persons.

representing or reconstructing the full information as included in each data block (which implies that we will, e.g., leave aside simple voting procedures and procedures of meta-analysis, or, more in general, high-level data fusion [13]). Further, regarding information to be derived from the whole of the data blocks, we will focus on problems that imply a primary research interest in the commonalities and differences between the different data blocks under study as well as in the linking relations between those blocks (which implies that we will leave aside methods that, e.g., exclusively focus on linking relations such as canonical correlation analysis methods). In terms of the second meta-question, we will exclusively focus on problems that imply the assumption of exchangeability of the different data blocks, both with regard to role and with regard to level of priority or importance (which implies that we will disregard regression-type of models). Complicating factors will further not be the primary focus of the present paper, although we will briefly touch upon a few of them when discussing research challenges.

In this section, we will deal with a family of methods that addresses the subclass of data fusion problems as outlined above. More in particular, we will propose a novel generic modeling framework for this family. This modeling framework will subsume a broad range of specific models (both existing and to be developed ones) as special cases. Below we will first give a formal definition of the generic model. Next we will give examples of existing methods that are subsumed by it. As an important special case, we will show how multiway models can be reconceived as models of data fusion subsumed by our generic model. We will conclude this section with a long list of research challenges that go with our generic modeling approach.

4.1. Formal definition

We assume a coupled data set, **D**, that comprises K linked data blocks ($\mathbf{B}_1$, …, $\mathbf{B}_k$, …, $\mathbf{B}_K$) (with block $\mathbf{B}_k$ involving N_k modes). The generic model is a global model for the whole of all K data blocks. This global model consists of: (1) a submodel for each data block that accounts for the individual data entries in that block, and (2) a linking structure between these submodels. We will now successively discuss each of those two aspects more in detail.

4.1.1. Submodel per data block

For the time being, we focus on a single data block **B**. We assume this constitutes an $(I_1 \times \ldots \times I_n \times \ldots \times I_N)$ N-way N-mode array. (An extension to the N-way N'-mode case is rather straightforward and will be briefly touched upon below.) The submodel for data block **B** is subsumed by a unifying model as proposed by Ref. [14]. The heart of this unifying model is deterministic in nature; yet, optionally, the deterministic heart can be extended with a stochastic error model to represent discrepancies between the actual entries in the data and the corresponding reconstructed entries in the deterministic heart of the model (for one possible general procedure to build a stochastic extension of a deterministic model, see Ref. [15]).

The unifying model as introduced in Ref. [14] comprises two ingredients: a quantification of each of the modes as involved in the data block, and an association rule that allows to reconstruct each entry of the data block on the basis of its implied mode-specific quantifications. Below we will successively discuss each of those two ingredients more in detail. Throughout, we will use a hypothetical gene by tissue transcriptomics two-way two-mode data block as a guiding example.

4.1.1.1. Quantifications of data block modes. The first constituent of the unifying model is a quantification of each of the N modes as involved in the data. This quantification can be considered a reduction of the mode in question. For the n-th data mode (which is assumed to comprise I_n elements), this quantification can be captured by means of an $I_n \times P_n$ matrix $\mathbf{A}^n$, the entries of which take either 0/1 or unconstrained real values. In the gene by tissue example, this would imply an $I_1 \times P_1$ quantification matrix $\mathbf{A}^1$ of the genes and an $I_2 \times P_2$ quantification matrix $\mathbf{A}^2$ of the tissues. In case a quantification matrix $\mathbf{A}^n$ would be purely 0/1, it would imply a clustering of the elements of the corresponding mode into P_n clusters. If there are no further constraints on the binary quantification (or cluster membership) matrix $\mathbf{A}^n$, the clustering of the corresponding mode would be an unconstrained overlapping one. Special cases of constrained binary quantification matrices include partitionings and nested clusterings. In case a quantification matrix $\mathbf{A}^n$ would be purely real-valued, it would imply a reduction or representation of the elements of the corresponding n-th data mode as points in a low-dimensional (i.e., a P_n-dimensional) space.

Four remarks can further be made regarding the mode-specific quantifications. First, for a single mode, the quantification can be either pure (i.e., purely binary or purely real-valued) or hybrid (i.e., mixed categorical–dimensional). Second, the nature of the quantification can differ across modes (e.g., one could go for a partitioning of the genes along with a low-dimensional representation of the tissues), which would yield another kind of hybrid (mixed categorical–dimensional) model. Third, we allow for the extreme or degenerate case of a quantification matrix $\mathbf{A}^n$ being an identity matrix, which would imply that the n-th data mode is not being reduced. Fourth, if the data would be N-way N'-mode rather than N-way N-mode (with $N' < N$), the quantification of the n-th mode is assumed to be the same across all ways pertaining to that mode. As an example, a model for two-way one-mode gene by gene similarity data would imply a single quantification matrix for the gene mode.

4.1.1.2. Block-specific association rule. In addition to a quantification of each of the N data modes involved in the block under study, with the $I_n \times P_n$ matrix $\mathbf{A}^n$ capturing the quantification of the n-th mode, the block-specific model includes a block-specific association rule. This rule includes a $(P_1 \times \ldots \times P_n \times \ldots \times P_N)$ core array **W** and a mapping f, which are such that:

$$\mathbf{B} = f\left(\mathbf{A}^1, \ldots, \mathbf{A}^N, \mathbf{W}\right) + \mathbf{E}, \tag{1}$$

with **E** denoting an $(I_1 \times \ldots \times I_n \times \ldots \times I_N)$ array with residuals or error entries, and with, from the point of view of the n-th mode ($n = 1, \ldots, N$), $f(\mathbf{A}^1, \ldots, \mathbf{A}^N, \mathbf{W})_{i_1 \ldots i_n \ldots i_N}$ depending only on the i_n-th row of $\mathbf{A}^n$. One may note that the latter means that for each data mode it holds that, for each element of that mode, all distinctive information on that element is contained in its corresponding row in the mode-specific quantification matrix (which means that this matrix does represent a reduction of the mode in question, indeed).

To clarify the quite broad concept of an association rule, we will illustrate this with a few specific examples. A first specific association rule is that of a generalized Cartesian product [16]:

$$f\left(\mathbf{A}^1, \ldots, \mathbf{A}^N, \mathbf{W}\right)_{i_1 \ldots i_n \ldots i_N} = \sum_{p_1=1}^{P_1} \ldots \sum_{p_n=1}^{P_n} \ldots \sum_{p_N=1}^{P_N} a^1_{i_1 p_1} \ldots a^n_{i_n p_n} \ldots a^N_{i_N p_N} w_{p_1 \ldots p_n \ldots p_N}. \tag{2}$$

In the two-way two-mode case, Expression (2) reduces to:

$$f\left(\mathbf{A}^1, \mathbf{A}^2, \mathbf{W}\right)_{i_1 i_2} = \sum_{p_1=1}^{P_1} \sum_{p_2=1}^{P_2} a^1_{i_1 p_1} a^2_{i_2 p_2} w_{p_1 p_2}, \tag{3}$$

or, in matrix form,

$$f\left(\mathbf{A}^1, \mathbf{A}^2, \mathbf{W}\right) = \mathbf{A}^1 \mathbf{W} \left(\mathbf{A}^2\right)^T. \tag{4}$$

In case the quantification matrices $\mathbf{A}^1$ and $\mathbf{A}^2$ are restricted to take 0/1 values, one arrives at a family of biclustering models ([17,18]), which within systems biology has been widely used to model gene by tissue transcriptomics data.

In case all quantification matrices are allowed to take real values, Expression (2) denotes the family of Tucker$_N$ models for multiway data, with in the more familiar three-way case:

$$f\left(\mathbf{A}^1,\mathbf{A}^2,\mathbf{A}^3,\mathbf{W}\right)_{i_1i_2i_3} = \sum_{p_1=1}^{P_1}\sum_{p_2=1}^{P_2}\sum_{p_3=1}^{P_3} a^1_{i_1p_1}a^2_{i_2p_2}a^3_{i_3p_3}w_{p_1p_2p_3}. \quad (5)$$

A systems biology application of this model can be found in Ref. [10].

A different branch of block-specific models is obtained when the mapping f in Eq. (1) involves a distance-type of construct. As an example, in the two-way case, one could consider the model

$$f\left(\mathbf{A}^1,\mathbf{A}^2,\mathbf{W}\right)_{i_1i_2} = \left[\sum_{p_1=1}^{P_1}\sum_{p_2=1}^{P_2}\left(a^1_{i_1p_1}-a^2_{i_2p_2}\right)^2 w_{p_1p_2}\right]^{\frac{1}{2}}. \quad (6)$$

In case $\mathbf{W}$ is an identity matrix, Eq. (6) reduces to:

$$f(\mathbf{A}^1,\mathbf{A}^2)_{i_1i_2} = \left[\sum_{p=1}^{P}\left(a^1_{i_1p}-a^2_{i_2p}\right)^2\right]^{\frac{1}{2}}. \quad (7)$$

This formalizes a model that in the psychological literature is denoted by the name multidimensional unfolding. A custom-made algorithm to fit a multidimensional unfolding model to gene by tissue transcriptomics data (called Genefold) has been developed by Ref. [19].

Other association rules are discussed in Ref. [14].

4.1.2. Linking structure between different submodels

We assume that the data set under study comprises K linked data blocks. We denote these further by ($\mathbf{B}_1$, …, $\mathbf{B}_k$, …, $\mathbf{B}_K$), with block $\mathbf{B}_k$ involving N_k modes. For each data block we further assume a block-specific submodel, that is:

$$\begin{aligned}
&\mathbf{B}_1 = f_1\left(\mathbf{A}^1_1,\ldots,\mathbf{A}^{N_1}_1,\mathbf{W}_1\right)+\mathbf{E}_1\\
&\ldots\\
&\mathbf{B}_k = f_k\left(\mathbf{A}^1_k,\ldots,\mathbf{A}^{N_k}_k,\mathbf{W}_k\right)+\mathbf{E}_k\\
&\ldots\\
&\mathbf{B}_K = f_K\left(\mathbf{A}^1_K,\ldots,\mathbf{A}^{N_K}_K,\mathbf{W}_K\right)+\mathbf{E}_K.
\end{aligned} \quad (8)$$

Note that in Eq. (8) as well as throughout this paper, subscripts of matrices, arrays, and functions will pertain to blocks, and superscripts of matrices and arrays to modes within blocks. Note further that in Eq. (8), the quantification matrices $\mathbf{A}$, the core array $\mathbf{W}$, and the linking functions f all bear a block-specific subscript, because all of them may in principle vary across blocks.

The fact that data blocks ($\mathbf{B}_1$, …, $\mathbf{B}_k$, …, $\mathbf{B}_K$) constitute coupled data simply means that they share a number of modes. In our global generic model, this mode sharing is captured through constraints on the quantification matrices of the shared modes. These constraints can be conceived as representing the linking structure of the model.

In principle, a broad range of constraints could be considered for the representation of linking structures. The most simple of them is an identity constraint. Such a constraint simply implies that a shared mode is given the same quantification in all submodels in which it shows up. To clarify this, we illustrate with two examples. As a first illustration, consider the part of the data of Guiding example 2 that is schematically represented in panel (b) of Fig. 2. This data part consist of three coupled two-mode blocks that all share a common mode (the genes). Without loss of generality, we can assume that for each data block, the gene set constitutes the first mode. A global model with an identity link for these data then would read as follows:

$$\begin{aligned}
&\mathbf{B}_1 = f_1\left(\mathbf{A}^1,\mathbf{A}^2_1,\mathbf{W}_1\right)+\mathbf{E}_1\\
&\mathbf{B}_2 = f_2\left(\mathbf{A}^1,\mathbf{A}^2_2,\mathbf{W}_2\right)+\mathbf{E}_2\\
&\mathbf{B}_3 = f_3\left(\mathbf{A}^1,\mathbf{A}^2_3,\mathbf{W}_3\right)+\mathbf{E}_3,
\end{aligned} \quad (9)$$

with the identity constraint being hidden in the subtle fact that the quantification matrix $\mathbf{A}^1$ does no longer bear a block-specific subscript.

As a second illustration, we consider the data of Guiding example 3 as schematically represented in panel (c) of Fig. 2. The data now consist of a two-mode and a three-mode block that are coupled through a single mode (pertaining to the genes). Without loss of generality, we again assume that for each data block, the gene set constitutes the first mode. A global model with an identity link for these data then would read as follows:

$$\begin{aligned}
&\mathbf{B}_1 = f_1\left(\mathbf{A}^1,\mathbf{A}^2_1,\mathbf{W}_1\right)+\mathbf{E}_1\\
&\mathbf{B}_2 = f_2\left(\mathbf{A}^1,\mathbf{A}^2_2,\mathbf{A}^3_2,\mathbf{W}_2\right)+\mathbf{E}_2,
\end{aligned} \quad (10)$$

with the identity constraint again being hidden in the subtle fact that the quantification matrix $\mathbf{A}^1$ does not bear a block-specific subscript.

Beyond a pure identity constraint, one may consider various other linking structures. Here we will list only a few possibilities.

Rather than a full identity constraint on the quantification matrices of shared modes, one could consider partial identity constraints. At this point, two forms of a partial identity constraint deserve a special mentioning. The first of these is that a number of columns of the quantification matrices of a shared mode are constrained to be identical, whereas other columns are left unconstrained. Through such a partial identity constraint, one may wish to capture both commonalities in the structures of the linked data blocks (in terms of the identical quantification columns) and distinctive aspects (in terms of the unconstrained columns). A second partial identity constraint reads that quantification matrices of a shared mode are constrained to be identical with regard to the vast majority of their rows. This means that for the vast majority of the elements of the mode involved, but not for all, the quantifications have to be the same (with elements that require different quantifications having to be identified during the data-analytic process). As an example, one may consider the coupled transcriptomics data pertaining to two different organisms as schematically represented in panel (a) of Fig. 4. In cases like this, a partial identity constraint may be called for, with different quantifications being needed for genes that underwent changes throughout evolutionary history. One may note that this second form of partial identity constraint bears a close relation to problems of (configural) measurement invariance that have been studied fairly extensively within psychometrics (see, e.g., Ref. [20]). We will come back to this issue below in Section 4.4.5.

Special types of linking structures may be needed if one of the shared modes is the time mode. In such cases, indeed, the linking structure may have to account for lags in dynamics. This is, for instance, the case if measurements of metabolites are performed in blood and urine, where usually the metabolite appears earlier in the blood.

We conclude this section by mentioning two other possible linking structures, both of which are asymmetric in nature. The first of these pertains to the case of binary quantification matrices (which can be conceived as membership matrices in some clustering). A constraint on such matrices could read that the clustering as implied by the first

quantification matrix is nested in the second. As a special case of this, in case one would consider partitioning matrices only, a nestedness constraint would imply that the first partitioning is a refinement of the second (i.e., the first partitioning then is to be obtained by splitting a number of classes of the second one). As a second possibility, in case of real-valued quantification matrices, one may require two quantifications of the same mode to be in a space–subspace relation.

4.2. Examples of instantiations of the generic linked-mode decomposition model

In chemometrics, SUM-PCA (or Consensus-PCA) is a much used data fusion method [4]. Related methods in psychometrics are multiple factor analysis (MFA; [21]) and STATIS [22]. The exact relationships between these methods have been published elsewhere [23] and will not be repeated here. All methods fit within our generic framework. This will be illustrated for the two-block case. Assuming that appropriate preprocessing has taken place per data block, SUM-PCA assumes that:

$$\begin{aligned}\theta_{m1}\mathbf{B}_1 &= \mathbf{A}^1\left(\mathbf{A}_1^2\right)^T + \mathbf{E}_1 = f\left(\mathbf{A}^1, \mathbf{A}_1^2\right) + \mathbf{E}_1\\ \theta_{m2}\mathbf{B}_2 &= \mathbf{A}^1\left(\mathbf{A}_2^2\right)^T + \mathbf{E}_2 = f\left(\mathbf{A}^1, \mathbf{A}_2^2\right) + \mathbf{E}_2,\end{aligned} \tag{11}$$

where, depending on the specific method, different weights θ_{mk} are assigned to the data blocks (see Ref. [23] for details). This is an example of a data fusion model with an identity link where the blocks $\mathbf{B}_1$ and $\mathbf{B}_2$ have the sampling mode in common. It can easily be generalized to more than two blocks of data. An example of the use of this method can be found in Ref. [24], where metabolomics and gene expression data are coupled in a toxicology experiment.

Whereas SUM-PCA pertains to two-way data sets, instantiations of the generic linked-mode decomposition model also exist for two coupled two-way and three-way data blocks that have a single mode in common [25]. An example of combining a PARAFAC and a PCA model for such data reads:

$$\begin{aligned}\mathbf{B}_1 &= \mathbf{A}^1\left(\mathbf{A}_1^2\right)^T + \mathbf{E}_1 = f_1\left(\mathbf{A}^1, \mathbf{A}_1^2\right) + \mathbf{E}_1\\ \mathbf{B}_2 &= \mathbf{A}^1\left(\mathbf{A}_2^2 \odot \mathbf{A}_2^3\right)^T + \mathbf{E}_2 = f_2\left(\mathbf{A}^1, \mathbf{A}_2^2, \mathbf{A}_2^3\right) + \mathbf{E}_2,\end{aligned} \tag{12}$$

where $\mathbf{A}_2^2$ and $\mathbf{A}_2^3$ are loading matrices pertaining to the second and third modes of the properly matricized three-way array $\mathbf{B}_2$, and $\odot$ is the symbol for the Khatri-Rao product [26].

A final example of a method subsumed by the generic model is linked-mode PARAFAC, as proposed by Harshman in Ref. [27] for coupled three-mode data blocks. For the special case of three three-mode data blocks, the first two of which having their first mode in common and the latter two their second mode, the model equations read as follows:

$$\begin{aligned}\mathbf{B}_1 &= \mathbf{A}^1\left(\mathbf{A}_1^2 \odot \mathbf{A}_1^3\right)^T + \mathbf{E}_1 = f\left(\mathbf{A}^1, \mathbf{A}_1^2, \mathbf{A}_1^3\right) + \mathbf{E}_1,\\ \mathbf{B}_2 &= \mathbf{A}^1\left(\mathbf{A}^2 \odot \mathbf{A}_2^3\right)^T + \mathbf{E}_2 = f\left(\mathbf{A}^1, \mathbf{A}^2, \mathbf{A}_2^3\right) + \mathbf{E}_2,\\ \mathbf{B}_3 &= \mathbf{A}_3^1\left(\mathbf{A}^2 \odot \mathbf{A}_3^3\right)^T + \mathbf{E}_3 = f\left(\mathbf{A}_3^1, \mathbf{A}^2, \mathbf{A}_3^3\right) + \mathbf{E}_3.\end{aligned} \tag{13}$$

4.3. Reconceiving multiway models as models of data fusion

As an interesting aside, one may note that single-block multiway data may be reconceived as multiblock data; reconceiving the associated multiway models along the same lines makes clear that in most cases such models can be considered as instantiations of our generic model.

To clarify this, let us focus on the three-way case. Starting from the slice mode, any single three-way three-mode $I\times J\times K$ data array $\mathbf{B}$ can be reconceived as a collection of K linked $I\times J$ data matrices ($\mathbf{B}_1$, …, $\mathbf{B}_k$, …, $\mathbf{B}_K$), with each matrix pertaining to one of the slices of the array $\mathbf{B}$. The matrices ($\mathbf{B}_1$, …, $\mathbf{B}_k$, …, $\mathbf{B}_K$) further all share the same row and column mode. Similar reconceptualizations are possible by starting from the row mode of the array $\mathbf{B}$, which would yield I linked $J\times K$ data matrices ($\tilde{\mathbf{B}}_1$, …, $\tilde{\mathbf{B}}_i$, …, $\tilde{\mathbf{B}}_I$), or from the column mode of the array $\mathbf{B}$, which would yield J linked $I\times K$ data matrices $\left(\tilde{\tilde{\mathbf{B}}}_1, \ldots, \tilde{\tilde{\mathbf{B}}}_j, \ldots, \tilde{\tilde{\mathbf{B}}}_J\right)$.

To understand the associated reconceptualization of three-way models, let us consider a PARAFAC model for $\mathbf{B}$. With the notation of Eq. (12), this reads as follows:

$$\mathbf{B} = \mathbf{A}^1\left(\mathbf{A}^2 \odot \mathbf{A}^3\right)^T + \mathbf{E} = f\left(\mathbf{A}^1, \mathbf{A}^2, \mathbf{A}^3\right) + \mathbf{E}, \tag{14}$$

or, with the notation of Eq. (5):

$$b_{ijk} = \sum_{p=1}^{P} a_{ip}^1 a_{jp}^2 a_{kp}^3 + e_{ijk}. \tag{15}$$

Taking into account the rewriting of the three-way array $\mathbf{B}$ as a collection of K linked matrices ($\mathbf{B}_1, \ldots, \mathbf{B}_k, \ldots, \mathbf{B}_K$), Eq. (15) can be rewritten as:

$$\begin{aligned}&(b_1)_{ij} = \sum_{p=1}^{P} a_{ip}^1 a_{jp}^2 a_{1p}^3 + (e_1)_{ij}\\ &\ldots\\ &(b_k)_{ij} = \sum_{p=1}^{P} a_{ip}^1 a_{jp}^2 a_{kp}^3 + (e_k)_{ij}\\ &\ldots\\ &(b_K)_{ij} = \sum_{p=1}^{P} a_{ip}^1 a_{jp}^2 a_{Kp}^3 + (e_K)_{ij}.\end{aligned} \tag{16}$$

In matrix notation, this then becomes:

$$\begin{aligned}&\mathbf{B}_1 = \mathbf{A}^1\mathbf{W}_1\left(\mathbf{A}^2\right)^T + \mathbf{E}_1 = f\left(\mathbf{A}^1, \mathbf{A}^2, \mathbf{W}_1\right) + \mathbf{E}_1\\ &\ldots\\ &\mathbf{B}_k = \mathbf{A}^1\mathbf{W}_k\left(\mathbf{A}^2\right)^T + \mathbf{E}_k = f\left(\mathbf{A}^1, \mathbf{A}^2, \mathbf{W}_k\right) + \mathbf{E}_k\\ &\ldots\\ &\mathbf{B}_K = \mathbf{A}^1\mathbf{W}_K\left(\mathbf{A}^2\right)^T + \mathbf{E}_K = f\left(\mathbf{A}^1, \mathbf{A}^2, \mathbf{W}_K\right) + \mathbf{E}_K\end{aligned} \tag{17}$$

with $\mathbf{W}_k (k=1, \ldots, K)$ being a $P\times P$ diagonal matrix with entries a_{kp}^3. Due to the symmetry of the PARAFAC model, similar models can be written for ($\tilde{\mathbf{B}}_1$, …, $\tilde{\mathbf{B}}_i$, …, $\tilde{\mathbf{B}}_I$) and $\left(\tilde{\tilde{\mathbf{B}}}_1, \ldots, \tilde{\tilde{\mathbf{B}}}_j, \ldots, \tilde{\tilde{\mathbf{B}}}_J\right)$. Similar reconceptualizations can be given for the Tucker3, Tucker2, and Tucker1 models.

4.4. Research challenges

The generic data model for data fusion as outlined above goes with a broad range of research challenges. Below we will briefly list a number of these, without making an attempt to be exhaustive. Successively we will discuss challenges on the level of the design for the data collection (Section 4.4.1), the actual modeling (Section 4.4.2), objective functions to be optimized during the data analysis (Section 4.4.3), algorithms (Section 4.4.4), and various data-analytic issues (Section 4.4.5).

4.4.1. Design for the data collection

In some cases one may wish to measure a set of batches, tissues, or organisms with regard to different sets of variables, without the possibility of measuring all batches etc. with regard to all variables. This then results in coupled data with a partially shared variable mode, as schematically represented in Fig. 5. In such cases, one may nevertheless wish to capture the structure of the batches (resp. tissues

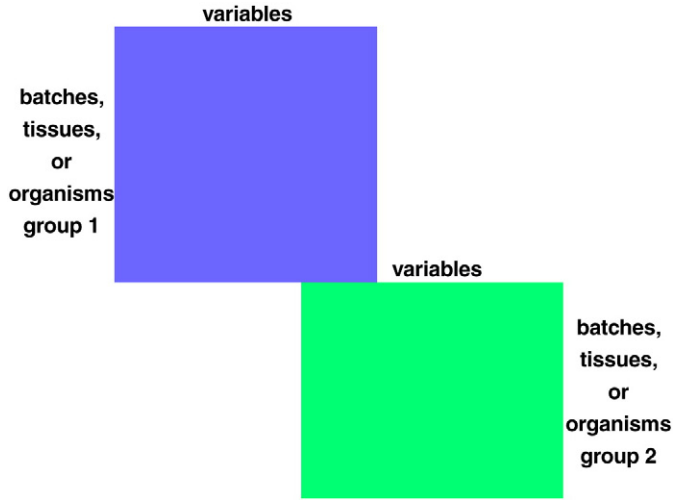

Fig. 5. Schematic representation of two coupled data blocks with partially shared variable mode.

or organisms) in terms of quantifications that are comparable across groups of batches. This problem has been studied extensively in psychometrics under the name 'test equating'. In the case of test equating, the variables in Fig. 5 pertain to items as grouped in different tests, and the different groups of batches to groups of respondents. A critical issue in problems of test equating is the specification of an appropriate design for the data collection that allows to arrive at comparable quantifications (e.g., ability estimates) for all groups of test takers or respondents. Such a design may, for instance, imply the use of a suitable overlap between the different tests (in terms of some kind of so-called anchoring items). Within a systems biology context, this would imply the specification of an appropriate subset of variables that are to be measured in all batches (tissues or organisms) under study. Consider as an example a case in which data are available on the concentration levels of certain metabolites in both a tissue (e.g., muscular tissue or adipose tissue) and a body fluid (e.g., blood or urine). A critical design issue then could pertain to the specification of the subset of metabolites that are to be measured in both the tissue and the body fluid.

A special challenge pertains to designing the fusing problem in such a way that the resulting data permit testing or exploring alternative linking structures. This is a completely unexplored area of research.

4.4.2. Model

The generic model for data fusion as introduced in the present paper subsumes a broad range of specific models (continuous, discrete as well as hybrid ones) as special cases. Challenges on the level of modeling first include continuing the endeavor we started in Section 4.2 to find out which existing models fit under our generic model and how. This may ultimately also lead to a better understanding of model interrelations and to building bridges between modeling traditions in different research disciplines.

A second challenge pertains to the specification of novel instantiations of the generic model that meet domain-specific needs (e.g., within the domain of systems biology). A first possible direction for such specifications could pertain to the development of various kinds of constraints that could be put on the mode-specific quantification matrices $\mathbf{A}_k^n$, on the linking arrays $\mathbf{W}_k$, or on the association rules f_k (for a typology of different kinds of constraints, see Ref. [28]). Otherwise, within a data fusion context, one may consider both block-specific constraints and constraints overarching multiple blocks; as an example of the latter, one might constrain the linking arrays $\mathbf{W}_k$ in Eq. (8) to be the same for all blocks. A second direction for model specification could pertain to the development of novel, custom-made linking structures between the different submodels of a data fusion representation.

Other challenges on the level of modeling include the study of uniqueness. For instance, many component models (including simultaneous and multiway component models) as subsumed by our generic model, are well known to be subject to rotational freedom. This and other types of nonuniqueness do not provide a basis for dismissing the models in question. Rather, they should be carefully studied and understood in order to arrive at correct interpretations of modeling results.

4.4.3. Objective function

Instantiations of our generic data fusion model can be either deterministic or stochastic in nature, depending on whether the submodels for the distinct data blocks do not or do include a stochastic model for the error terms (see Section 4.1.1). As a consequence, the objective or loss functions that is to be optimized in the data analysis can be based on a direct measure of the discrepancies between actual and reconstructed data entries (of, e.g., a least squares or, more in general, a least L_p-norm type), or can be based on a likelihood or posterior distribution function. Typically, objective functions that go with data fusion models are single criterion functions that are compounds of objective subfunctions pertaining to the submodels for the distinct data blocks.

Many challenges with regard to the objective function relate to looking for optimal combinations of the block-specific objective subfunctions, optimality to be understood here as leading to optimal inferences (about the true structures underlying the data blocks etc.). One subgroup of challenges at this point pertains to the operator through which the subfunctions are to be combined; examples of such operators include addition and multiplication, the latter operator possibly leading to better representations of commonalities between the different data blocks (see Ref. [29]). Another subgroup of challenges pertains to dealing with possible differences between the different data blocks in terms of size and in terms of noise level, as already addressed in Section 3.3. Given some operator to aggregate the different block-specific subfunctions (e.g., addition), one may consider to deal with possible between-block differences by means of the inclusion in the aggregation of suitable block-specific weights. An example of a search of suitable weights to deal with between-block differences in size within the context of a specific data fusion model can be found in Ref. [25]; for a study on the search of appropriate weights to deal with between-block differences in noise level (custom-made for a systems biology context), see Ref. [30].

4.4.4. Algorithms

An obvious challenge associated with the generic linked-mode decomposition model for data fusion is the development of suitable algorithmic strategies to optimize the objective functions as associated with specific instantiations of it. Subsequently, such strategies are to be evaluated in terms of their optimization performance (including an investigation into their susceptibility to local optima), and of their computational efficiency and feasibility (in particular with regard to data sets with sizes that typically occur in some target research domain, such as systems biology).

A fairly broad class of strategies that could be considered within the context of our generic data fusion model is that of the iterative alternating type. Such strategies presume a partitioning of the full set of model parameters. After an initialization, each of the partition classes is further updated by optimizing the objective function, conditional upon the current values for the parameters of all other partition classes [31]. This updating is to be repeated until no further (sizeable) gain in the objective function can be obtained.

Leaving aside for a moment the parameters of the optional stochastic error models, an obvious partitioning of the parameters in the case of our generic data fusion model of Eq. (8) is that into the different mode-specific quantifications and block-specific linking arrays ($\mathbf{A}_1^1$, …, $\mathbf{A}_K^{N_K}$, $\mathbf{W}_1$, …$\mathbf{W}_K$). In such a partitioning, one can draw a distinction between partition classes of parameters that are not and that are subject to constraints involved in the linking structure between the coupled blocks. A conditional updating of a partition class that is not involved in such block-overarching constraints can be done taking into account only the specific data block to which that partition class pertains. A conditional updating of a partition class that is subject to some linking constraint, however, is more involving, as it necessarily is to be based on all blocks to which the linking constraint in question pertains.

4.4.5. Data-analytic issues

The generic linked-mode decomposition model for data fusion as introduced in the present paper goes with a wealth of data-analytic challenges. An important subclass of those challenges pertains to model selection issues. Below, we will treat this in separate subsection. Subsequently, we will briefly discuss a few other data-analytic issues.

4.4.5.1. Model selection. Our generic model has a very broad scope. This broad scope is associated with a long list of choices that are to be made with regard to many kinds of modeling aspects or options. Part of these choices will typically have to be made on an a priori basis, that is, on the basis of domain-specific theoretical concerns and a priori preferences of the researcher. Another part of the choices could be data-driven; this implies the challenge of developing suitable model selection procedures and heuristics (which subsequently are to be evaluated on theoretical and empirical grounds).

Below, we will draw a distinction between three groups of model selection issues. We will successively discuss each of those three.

(1) A first group of model selection issues pertains to structural aspects of the data as they will be dealt with during the data-analytic process. These aspects include (1a) the block structure and (1b) the linking structure between the blocks.
   1a. On the block level, an important question reads which elements should be kept together as a single mode in one of the data blocks. As an example, one may refer to the data of Guiding example 1 as schematically represented in panel (a) of Fig. 2. A critical decision for these data pertains to whether the metabolites as measured by the two measurement platforms (LC/MS and GC/MS) are to be considered as a single mode (resulting in a single data block) or rather as two distinct modes (resulting in two coupled data blocks). It is important to emphasize that such mode-splitting decisions may be fairly consequential for the results of a subsequent modeling, for instance, if such a modeling would include some type of block-based preprocessing.
   1b. On the level of the linking structure, global as well as local choices are to be made:

   - Global linkage choices: Those pertain to whether modes as a whole are to be shared between different data blocks. As an example, we reconsider the case of a PARAFAC model for three-way three-mode data as formalized by Eq. (17). This equation implies that the data are conceived as K data blocks (matrices), all of which share the same row and column mode. One may, however, wonder whether such an assumption makes sense from a content-related point of view. As an example, one may think of gene by time point by tissue data as graphically represented in panel (b) of Fig. 1. Linking up with the argument of Section 4.3, such data may be reconceived as K gene by time point matrices (one for each tissue), which are linked via both the gene and the time point mode. Yet, from a content-related point of view, one could possibly have some concern about sharing a single quantification of time among all tissues. Perhaps, between-tissue differences in gene expression time courses could be better captured by a tissue-specific quantification of time. This would imply replacing Eq. (17) by the following:

$$\begin{aligned}
&\mathbf{B}_1 = \mathbf{A}^1\mathbf{W}_1\left(\mathbf{A}_1^2\right)^T + \mathbf{E}_1 = \mathbf{A}^1\left(\tilde{\mathbf{A}}_1^2\right)^T + \mathbf{E}_1\\
&\ldots\\
&\mathbf{B}_k = \mathbf{A}^1\mathbf{W}_k\left(\mathbf{A}_k^2\right)^T + \mathbf{E}_k = \mathbf{A}^1\left(\tilde{\mathbf{A}}_k^2\right)^T + \mathbf{E}_k\\
&\ldots\\
&\mathbf{B}_K = \mathbf{A}^1\mathbf{W}_K\left(\mathbf{A}_K^2\right)^T + \mathbf{E}_K = \mathbf{A}^1\left(\tilde{\mathbf{A}}_K^2\right)^T + \mathbf{E}_K,
\end{aligned} \tag{18}$$

     which happens to be the model equation of the so-called PCA-SUP (or Tucker1) model as discussed in Ref. [32]. Formally speaking, the transition from Eq. (17) to Eq. (18) comes down to an undoing of the coupling of the time point mode across all tissues.

   - Local linkage choices: Those pertain to the question whether certain linkage constraints should hold for all elements of a shared mode or rather for some part of it. As an example, one may prefer an identity linkage constraint to hold for the vast majority of the elements of a shared mode, but not for all (as already referred to in Section 4.1.2). Within psychometrics, this issue has been studied extensively within the context of a fixed set of items that has been presented to several groups of respondents (e.g., stemming from different cultural backgrounds). This gives rise to respondent by item matrices (one per cultural group) that are linked through the item mode. In psychometrics, the model selection issue as to whether a data fusion model with a fully shared item quantification holds is denoted by the term (configural) measurement invariance. In this regard, it could be that for some items the assumption of a quantification that is shared among all cultural groups is to be rejected, a phenomenon that is known under the names of item bias or differential item functioning (DIF). This could pave the way for alternative models with a partial identity constraint on the item quantifications, in which the quantification of biased items is allowed to differ across cultural groups. Obviously, the psychometric work on item bias and DIF could be most relevant for studies in the area of comparative genomics (as illustrated in panel (a) of Fig. 4), DIF now to be translated into 'differential gene functioning'.

(2) A second group of model selection aspects pertains to the type of data fusion model that is to be chosen. This involves a specification of both the type of block-specific submodels and the nature of the linking structure. Regarding the type of submodels, choices to be made include: (a) the decision as to which modes are to be reduced, along with the type of reduction (continuous, discrete, hybrid), and (b) the choice of the decomposition function [14].
(3) A third group of model selection issues is more quantitative in nature, as it pertains to the extent of reduction for all data modes of all data blocks involved in the data fusion process (as

captured by the number of columns of the mode-specific quantification matrices $\mathbf{A}_k^n$). This group of model selection questions looks like the most straightforward or superficial one. This, however, does not mean that these questions are easy to solve, in particular because as a whole they imply a highly multivariate model selection problem.

4.4.5.2. Other data-analytic challenges. Beyond model selection, data fusion problems imply a broad range of other data-analytic challenges. Here we limit ourselves by mentioning only a few of them.

A first group of challenges (which rather closely relates to model selection) pertains to assessing goodness of fit on different levels (that of the coupled data as a whole, that of distinct data blocks, and that of distinct elements within a mode of some data block). A second group of challenges pertains to the identification and representation of estimation uncertainty. An example of a procedure to arrive at such an identification within a data fusion context can be found in Ref. [33]. Last but not least, there is the issue of arriving at sound interpretations of the results of a data fusion modeling endeavor. One challenging aspect in this regard pertains to the interpretation of quantifications of shared modes. One may expect such quantifications to contain information on both common aspects of all data blocks that share the mode in question, and distinctive aspects that pertain to a single block (or a few blocks) only. This implies the challenge of teasing apart the common and distinctive information as included in a quantification matrix of a shared mode at hand.

5. Practical relevance

We conclude this paper by showing how the proposed formal data description, meta-questions, and generic model may assist the data-analyst in choosing and developing suitable strategies for the treatment of coupled data in practice. Assume that transcriptomics data are available from tissues of different types (e.g., different types of cancer) along with protein binding information.

First, the formal description of coupled data as outlined in the present paper may provide some guidance in characterizing the type of the data. The transcriptomics information can be captured by means of a two-way two-mode gene by tissue matrix. Tissue type information can be represented by a two-way two-mode tissue by feature matrix, which shares the tissue mode with the transcriptomics data block. Finally, protein binding information can be captured by a two-way two-mode transcription factor by gene (binding site) matrix, which shares the gene mode with the transcriptomics data block. Taken together, this yields the data structure as schematically represented in Fig. 6.

Next, we can focus on the domain-specific research questions as associated with the data set at hand, making use of the meta-questions as outlined in this paper. First, we may wish to know whether the research interest resides in partial information that is derived from the coupled data under study, or that the systems biologist rather cares about reconstructing the full data set. If the primary interest would, for instance, be in detecting one or several modules to which some target genes belong, deriving from the coupled data blocks similarities with those target genes would be an appropriate choice. If, on the contrary, one would like to understand the full interplay of protein binding and gene expression, it may be more appropriate to try to capture the full information as included in the data. Second, we may wish to know whether, from the point of view of the systems biological researcher, the different data blocks are exchangeable in terms of role and importance. One possible outcome at this point could be that one may opt to predict gene expression on the basis of protein binding (which could perhaps further lead to the choice of some form of principal covariate regression analysis). As an alternative, one could go for an exchangeable scenario; in that case, one could look for suitable data-analytic methods subsumed by our generic linked-mode decomposition model for data fusion.

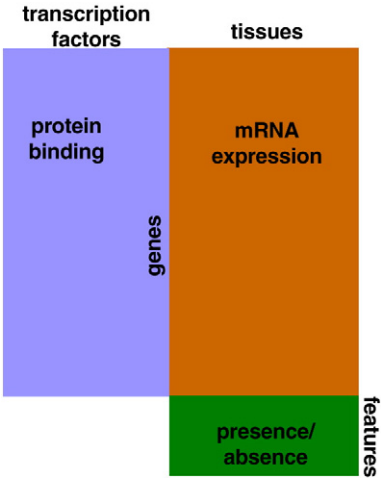

Fig. 6. Schematic representation of data structure in illustration of practical relevance of approach proposed in this paper.

If we go for a linked-mode decomposition model, quite a few other issues have to be decided upon, including the nature of the mode-specific quantifications and of the linking structure. For example, if the interest of the systems biologist would be in the detection of gene modules, one may opt for a binary quantification of the gene mode, which would result in a gene clustering. For the tissues, one might wish to end up with a clustering, too, perhaps with the linking constraint that this clustering is nested in the tissue types as represented by the tissue by feature matrix. Ultimately, this type of reasoning might pave the way for the creation of a novel, custom-made three-part linked-mode biclustering model for the data at hand.

Acknowledgements

The research reported in this paper was supported in part by the Research Fund of Katholieke Universiteit Leuven (EF/05/007 SymBioSys), by IWT-Flanders (IWT/060045/SBO Bioframe), and by Belgian Federal Science Policy (IAP/P6/03).

References

[1] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L. Tranchevent, B. De Moor, P. Marynen, B. Hassan, P. Carmeliet, Y. Moreau, Gene prioritization through genomic data fusion, Nature Biotechnology 24 (2006) 719.
[2] A. Smilde, M. van der Werf, S. Bijlsma, B. van der Werff-van-der Vat, R. Jellema, Fusion of mass spectrometry-based metabolomics data, Analytical Chemistry 77 (2005) 6729–6736.
[3] T. Wilderjans, E. Ceulemans, I. Van Mechelen, Simultaneous analysis of coupled data blocks differing in size: a comparison of two weighting schemes, Computational Statistics and Data Analysis 53 (2009) 1086–1098.
[4] A. Smilde, J. Westerhuis, S. de Jong, A framework for sequential multiblock component methods, Journal of Chemometrics 17 (2003) 323–337.
[5] P. Curran, A. Hussong, Integrative data analysis: the simultaneous analysis of multiple data sets, Psychological Methods 14 (2009) 81–100.
[6] D. Crockford, E. Holmes, J. Lindon, R. Plumb, S. Zirah, S. Bruce, P. Rainville, C. Stumpf, J. Nicholson, Statistical heterospectroscopy, an approach to the integrated analysis of NMR and UPLC–MS data sets: application in metabonomic toxicology studies, Analytical Chemistry 78 (2006) 363–371.
[7] O. Alter, P. Brown, D. Botstein, Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms, Proceedings of the National Academy of Sciences of the United States of America 100 (2003) 3351–3356.
[8] Y. Noguchi, Q. Zhang, T. Sugimoto, Y. Furuhata, Y. Sakai, M. Mori, M. Takahashi, T. Kimura, Network analysis of plasma and tissue amino acids and the generation of an amino index for potential diagnostic use, American Journal of Clinical Nutrition 83 (2006) 513S–519S.
[9] M. de Tayrac, S. Lê, M. Aubry, J. Mosser, F. Husson, Simultaneous analysis of distinct omics data sets with integration of biological knowledge: multiple factor analysis approach, BMC Genomics 10 (2009) 32.

[10] L. Omberg, G.H. Golub, O. Alter, A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies, Proceedings of the National Academy of Sciences of the United States of America 104 (2007) 18371–18376.
[11] L. Tucker, The extension of factor analysis to three-dimensional matrices, in: N. Frederiksen, H. Gulliksen (Eds.), Contributions to Mathematical Psychology, Erlbaum, Hillsdale, 1964, pp. 109–127.
[12] D. Rubin, Inference and missing data, Biometrika 63 (1976) 581–592.
[13] V. Steinmetz, F. Sevila, V. Bellon-Maurel, A methodology for sensor fusion design: application to fruit quality assessment, Journal of Agricultural Engineering Research 74 (1999) 21–31.
[14] I. Van Mechelen, J. Schepers, A unifying model involving a categorical and/or dimensional reduction for multimode data, Computational Statistics & Data Analysis 52 (2007) 537–549.
[15] A. Gelman, I. Leenen, I. Van Mechelen, P. De Boeck, J. Poblome, Bridges between deterministic and probabilistic models for binary data, Statistical. Methodology 7 (2010) 187–209.
[16] J. Carroll, A. Chaturvedi, A general approach to clustering and multidimensional scaling of two-way, three-way, or higher-way data, in: R. Luce, M. D'Zmura, D. Hoffman, G. Iverson, A. Romney (Eds.), Geometric Representations of Perceptual Phenomena, Erlbaum, Mahwah, 1995, pp. 295–318.
[17] I. Van Mechelen, H.-H. Bock, P. De Boeck, Two-mode clustering methods: a structural overview, Statistical Methods in Medical Research 13 (2004) 363–394.
[18] S. Madeira, A. Oliveira, Biclustering algorithms for biological data analysis: a survey, IEEE Transactions on Computational Biology and Bioinformatics 1 (2004) 24–25.
[19] K. Van Deun, K. Marchal, W. Heiser, K. Engelen, I. Van Mechelen, Joint mapping of genes and conditions via multidimensional unfolding analysis, BMC Bioinformatics 8 (2007) 181.
[20] W. Meredith, Measurement invariance, factor analysis and factorial invariance, Psychometrika 58 (1993) 525–543.
[21] J. Pagès, Collection and analysis of perceived product inter-distances using multiple factor analysis: application to the study of 10 white wines from the Loire valley, Food Quality and Preference 16 (2005) 642–649.
[22] H. L'Hermier des Plantes, B. Thiebaut, Etude de la pluviosite au moyen de la methode STATIS, Revue de Statistique Appliquée 25 (1977) 57–81.
[23] K. van Deun, A. Smilde, M. van der Werf, H. Kiers, I. van Mechelen, A structured overview of simultaneous component based data integration, BMC Bioinformatics 10 (2009) 246.
[24] W. Heijne, R. Lamers, P. van Bladeren, J. Groten, J. van Nesselrooij, B. van Ommen, Profiles of metabolites and gene expression in rats with chemically induced hepatic necrosis, Toxicology Pathology 33 (2005) 425–433.
[25] T. Wilderjans, E. Ceulemans, H. Kiers, K. Meers, The LMPCA program: a graphical user interface for fitting the linked-mode PARAFAC-PCA model to coupled real-valued data, Behavior Research Methods 41 (2009) 1073–1082.
[26] A.K. Smilde, R. Bro, P. Geladi, Multiway analysis. Applications in the Chemical Sciences, John Wiley & Sons, New York, 2004.
[27] R. Harshman, M. Lundy, PARAFAC: parallel factor analysis, Computational Statistics & Data Analysis 18 (1994) 39–72.
[28] E. Ceulemans, I. Van Mechelen, P. Kuppens, Adapting the formal to the substantive: constrained Tucker3-HICLAS, Journal of Classification 21 (2004) 19–50.
[29] H. Kiers, A. Smilde, A comparison of various methods for multivariate regression with highly collinear variables, Statistical Methods and Applications 16 (2007) 193–228.
[30] R. van den Berg, I. Van Mechelen, T. Wilderjans, K. Van Deun, H. Kiers, A. Smilde, Integrating functional genomics data using maximum likelihood based simultaneous component analysis, BMC Bioinformatics 10 (2009) 340.
[31] J. De Leeuw, Block-relaxation algorithms in statistics, in: H. Bock, W. Lenski, M. Richter (Eds.), Information Systems and Data Analysis, Springer Verlag, Berlin, 1994, pp. 308–325.
[32] H. Kiers, Hierarchical relations among three-way methods, Psychometrika 56 (1991) 449–470.
[33] M. Timmerman, H. Kiers, A. Smilde, E. Ceulemans, J. Stouten, Bootstrap confidence intervals in multilevel simultaneous component analysis, British Journal of Mathematical & Statistical Psychology 62 (2009) 299–318.

Glossary

association rule: Rule that indicates how an entry of a data block can be reconstructed on the basis of (a) the quantifications of the elements of all data modes to which this element pertains, and (b) the core array.
core array: Array for a data block at hand with weights for all possible combinations of clusters or dimensions involved in the mode-specific quantifications of that block.
coupled data: Connected whole of data blocks that are linked to one another through (fully or partially) shared block modes.
data block: Basic constituent of coupled data that can be formalized as a mapping from a Cartesian product of sets to some range (such as the real numbers or {0, 1}).
data fusion: Any type of data-analytic or statistical procedure that involves coupled data.
identity constraint: Constraint that reads that the block-specific quantifications of some data mode that is shared between several data blocks should be the same. This constraint constitutes the most basic example of a representation of a linking structure between data blocks.
linking structure: Constraints on block-specific quantifications of data modes that are shared between several blocks in coupled data, which formally represent the links between those blocks.
mode: Set of elements as involved in a data block; the number of modes of a data block equals the number of different sets involved in the Cartesian product underlying the block in question.
quantification: Reduction of the elements of a mode involved in some data block to memberships in a small number of clusters or/and to coordinates on a small number of dimensions.
way: Set of elements as included in the Cartesian product underlying a data block; the number of ways of a data block equals the total number of (possibly identical) sets involved in that Cartesian product.